A toolbox of permutation tests for structural change

Zeileis, Achim ; Hothorn, Torsten

DOI: https://doi.org/10.1007/S00362-013-0503-4

A Toolbox of Permutation Tests for Structural Change

Achim Zeileis
Universität Innsbruck

Torsten Hothorn
Universität Zürich

Abstract

The supLM test for structural change is embedded into a permutation test framework for a simple location model. The resulting conditional permutation distribution is compared to the usual (unconditional) asymptotic distribution, showing that the power of the test can be clearly improved in small samples. Furthermore, the permutation test is embedded into a general framework that encompasses tools for binary and multivariate dependent variables as well as model-based permutation testing for structural change. It is also demonstrated that the methods can not only be employed for analyzing structural changes in time series data but also for recursive partitioning of cross section data. The procedures suggested are illustrated using both artificial data and empirical applications (number of youth homicides, employment discrimination data, carbon flux in tropical forests, stock returns, and demand for economics journals).

Keywords: conditional inference, asymptotic distribution, exact distribution, maximally-selected statistics.

1. Introduction

Methods for detecting structural changes in series of observations have been receiving increased interest in the theoretical and applied literature, both in econometrics and statistics–see, e.g., Stock and Watson (1996) for a discussion of their relevance to econometric practice. Since the suggestion of the Quandt test (supremum of Chow statistics, see Chow 1960; Quandt 1960), several ideas for capturing structural instabilities in tests statistics have emerged. However, tracking the distribution of such test statistics (in the case of unknown timing of change) turned out to be difficult so that the (asymptotic) distribution of the Quandt test remained unknown for a long time. The breakthrough in deriving an asymptotic approximation for structural change test statistics came with the discovery of suitable functional central limit theorems, first for CUSUM statistics (Brown, Durbin, and Evans 1975, see also Krämer and Sonnberger 1986 for an early overview), then for sup$F$ statistics (Andrews 1993)—a unifying view on both types of tests is given in Zeileis (2005). Test procedures based on these asymptotic distributions are predominantly used in econometric practice, although some approaches for finite samples also employ other approximations, e.g., based on simulation or bootstrap sampling.

In this paper, we consider a different approach, namely conditional inference methods, also known as *permutation tests*. This powerful general principle for deriving a suitable reference distribution for a test statistic was described almost 80 years ago by Fisher (1935); its

asymptotic properties have been investigated early, e.g., by Pitman (1938). The approach has gained much popularity in the statistics literature in recent years (Ludbrook and Dudley 1998; Strasser and Weber 1999; Pesarin 2001; Ernst 2004). Permutation tests have been found particularly useful because of their flexibility, distribution-free nature and intuitive formulation, which makes it easy to communicate the general principles of such test procedures to practitioners.

While permutation tests have become increasingly popular in various statistics communities, in particular for nonparametric inference in biostatistics, they are still less popular in econometrics. However, several applications exist, see e.g., Kennedy (1995) for an overview on how to employ permutation tests in econometrics. For the specific problem of structural change, permutation tests are not typically considered in econometrics, whereas in the statistics literature they are used more frequently, especially in nonparametric tests based on maximally selected rank statistics (Lausen and Schumacher 1992; Hothorn and Lausen 2003; Boulesteix and Strobl 2007) and in mathematical statistics (Antoch and Hušková 2001; Kirch and Steinebach 2006; Kirch 2007). Building on these ideas, we discuss in the following how a wide class of permutation tests for structural change can be established, pointing out their strengths and weaknesses. The tests are derived within the framework of Strasser and Weber (1999) as discussed by Hothorn, Hornik, van de Wiel, and Zeileis (2006a), and using ideas of Kennedy (1995). In Section 2, the permutation distribution of the supLM test of Andrews (1993) is derived for the location-shift model and compared to the established (unconditional) asymptotic distribution, both for artificial data and in an empirical application. In Section 3, a general class of permutation tests for structural change is suggested and specific tests for binary and multivariate observations are derived as well as model-based permutation tests. It is demonstrated how the tests can not only be employed for assessing structural change in time series data but also for recursive partitioning of cross-sectional data. All procedures are illustrated using various empirical applications: the number of youth homicides in Boston, a case of employment discrimination, carbon flux due to coarse woody debris in tropical forests, Dow Jones industrial average stock returns, and the price elasticity of the demand for economics journals. Section 4 concludes the paper with a brief discussion.

2. Structural changes in the mean

For comparing unconditional and conditional inference techniques in a structural change context, we initially focus on the simple, yet important, special case of location shifts in a univariate series of observations. First, we establish some general notation as well as the general testing problem which is subsequently specialized to location shifts for which test statistics and sampling distributions are derived. Further important special cases of the general testing problem are considered in Section 3.

2.1. Test problem and statistics

Consider a sequence of n observations (or realizations) from random variables Y_i ($i = 1, \ldots, n$), possibly vector-valued, which is ordered with respect to t_i, usually corresponding to time (but could also be some other ordering variable such as income etc.). In the following, we assume that time t has been scaled to the unit interval such that it gives the fraction of observations up to the current time (without loss of generality). In the simplest case of n totally ordered

observations along $i = 1, \ldots, n$, this is simply $t_i = i/n$—in the more general case, there could be ties in t (i.e., several observations were made at the same time t).

The structure of the sequence Y_i is stable if the distribution of the observations $Y_i \sim \mathcal{F}_{t_i}$ does not depend on the time t_i. Thus, structural change tests are concerned with testing the hypothesis

$$H_0: \quad \mathcal{F}_t = \mathcal{F} \quad (t \in [0,1]) \tag{1}$$

against the alternative that the distribution $\mathcal{F}_t$ does depend on t in some way. As it is not possible (nor desired, typically) to derive tests that have good power properties under arbitrary alternatives (because there are infinitely many different ways how $\mathcal{F}_t$ can depend on t), specific test statistics are typically derived for certain patterns of deviation from the null hypothesis. The alternative most commonly of interest in this context is the single shift alternative, where the distribution remains constant up to an unknown breakpoint t^* and shifts to a different distribution afterwards. Test statistics derived for this particular alternative will, of course, also be able to pick up other structural changes albeit with less power. However, the loss in power is usually small if the true alternative can be described sufficiently well by a single shift.

Even a single shift alternative, however, is still too general if it is not specified which aspects of $\mathcal{F}$ are subject to change at t^*. To illustrate the basic approach and focus on the derivation of the conditional distribution of the test statistic, we consider in the remainder of this section the simplest case: only the first moment of $\mathcal{F}$ changes at t^*—more general types of changes in $\mathcal{F}$ are discussed in Section 3. In the case of location shifts, the model can be formulated more conveniently as

$$Y_i \quad = \quad \mu_{t_i} \; + \; \varepsilon_i, \tag{2}$$

where ε_i is a zero mean disturbance term. The null hypothesis and alternative can then be written as:

$$\begin{aligned} H_0: &\quad \mu_t = \mu \quad (t \in [0,1]) \\ H_A: &\quad \mu_t = \mu \quad (t \leq t^*) \\ &\quad \mu_t = \mu + \delta \quad (t > t^*) \end{aligned} \tag{3}$$

To test for single shift alternatives, the sup$F$ tests of Andrews (1993) are probably the tests employed most often in practice. In a mean shift model and using Lagrange multiplier (LM) statistics, the sup$F$ test is based on the statistics

$$F_\pi \quad = \quad n \cdot R_\pi^2 \; = \; n \cdot \left(1 - \frac{RSS_\pi}{RSS_0}\right), \tag{4}$$

where $RSS_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2$ is the residual sum of squares (RSS) under the null hypothesis and RSS_π is the RSS if the mean of the observations up to π is estimated by $\bar{Y}_{1,\pi}$ and the mean of the observations afterwards by $\bar{Y}_{2,\pi}$. A sequence of LM statistics F_π ($\pi \in \Pi$) is computed for each conceivable breakpoint $\pi \in \Pi = [\underline{\pi}, \overline{\pi}] \subset [0,1]$ and the overall null hypothesis is rejected if their supremum $\sup_{\pi \in \Pi} F_\pi$ is too large. The interval Π is typically derived using some trimming, e.g., $\Pi = [0.1, 0.9]$, which we use in our simulations and applications below.

For deriving the asymptotic conditional distribution in the following subsection, it will be useful to transform the LM statistics from their usual "F type" to a "t type", essentially by

taking the square root. The statistic F_π can be rewritten as follows

$$\begin{aligned} F_\pi &= \frac{RSS_0 - RSS_\pi}{RSS_0/n} \\ &= \frac{n_{1,\pi} n_{2,\pi}}{n} \frac{(\bar{Y}_{1,\pi} - \bar{Y}_{2,\pi})^2}{RSS_0/n} \end{aligned} \tag{5}$$

where $n_{1,\pi}$ and $n_{2,\pi}$ are the observations up to π and after π, respectively. Therefore, the supLM test can also be carried out by rejecting the null hypothesis if $\sup_{\pi\in\Pi} |Z_\pi|$ is too large, where

$$Z_\pi = \sqrt{\frac{n_{1,\pi} n_{2,\pi}}{n}} \frac{\bar{Y}_{1,\pi} - \bar{Y}_{2,\pi}}{\sqrt{RSS_0/(n-1)}}. \tag{6}$$

We have slightly rescaled Z_π by using $n-1$ instead of n for standardization. The reason for this is the derivation of the asymptotic conditional distribution and will be explained in more detail below. For notational convenience, we will sometimes replace the supremum by the maximum in the formulation of the test statistic $\max_{\pi\in\Pi} |Z_\pi|$—in these cases, Π is taken to be the elements of $[\underline{\pi}, \overline{\pi}]$ observed in the sample, i.e., $\Pi = \{t_i \,|\, \underline{\pi} \leq t_i \leq \overline{\pi}\} = \{\pi_1, \ldots, \pi_m\}$.

2.2. Distribution of the test statistic

In the previous subsection, two equivalent formulations of the supLM test have been established: reject the null hypothesis if $\max_{\pi\in\Pi} F_\pi$ or $\max_{\pi\in\Pi} |Z_\pi|$ becomes "too large". To render this test useful, the distribution $\mathcal{D}$ (or at least an approximation thereof) of the test statistic under the null hypothesis is required to compute critical values or equivalently p values. In general, unfortunately, the distribution $\mathcal{D} = \mathcal{D}_\mathcal{F}$ depends on the unknown distribution $\mathcal{F}$ and is therefore unknown as well. However, there are several strategies to dispose of this dependency by using a suitable approximation of $\mathcal{D}$. The most popular strategy in classical statistics and econometrics is to use the (unconditional) asymptotic distribution $\mathcal{D}_\infty$, i.e., to derive the limit of $\mathcal{D}_\mathcal{F}$ for $n \to \infty$ analytically under some (typically mild) regularity conditions.

In the case of the supLM test, this problem was solved in the seminal paper of Andrews (1993) who showed that a functional central limit theorem holds for the sequence of LM statistics F_π which converge to a squared standardized tied-down Bessel process under fairly general regularity conditions. Thus, the unconditional limiting distribution $\mathcal{D}_\infty$ is given by $\sup_{\pi\in\Pi}(\pi(1-\pi))^{-1}B^2(\pi)$, where $B(t)$ ($t \in [0,1]$) is a standard Brownian bridge. This distribution is nonstandard but efficient numerical algorithms for computing approximate p values from this distribution have been derived by Hansen (1997).

A fundamentally different strategy is to replace the unknown null distribution by the conditional null distribution, i.e., the distribution of the test statistic given the observed data. This approach leads to *permutation tests*, rendered computationally feasible by modern computers and therefore studied intensively today, see Ludbrook and Dudley (1998); Strasser and Weber (1999); Pesarin (2001); Ernst (2004), among others. In econometrics, the interest in permutation or randomization tests has also increased (see e.g., Kennedy 1995; Luger 2006) but less compared to the statistics community. An introduction to permutation tests in econometrics—highlighting both advantages and problems—is given by Kennedy (1995). Conceptually, carrying out a permutation test for structural change is easy: If the distribution of the Y_i does not depend on the time t_i, the Y_i can be permuted on the t_i, breaking up the

original ordering. The exact conditional distribution $\mathcal{D}_{\sigma|Y}$ of the test statistic can then be derived by computing the test statistic for each permutation $\sigma \in S$ of the observations Y_i. As the size of S is $n!$, it is only feasible for very small n to actually compute all permutations. Otherwise, either specialized algorithms are required for computing the exact distribution (which are only available in certain special cases) or it can be always approximated arbitrarily precisely by drawing a sufficiently large number of permutations P from S. In the following, we always draw $P = 10,000$ permutations to approximate the exact conditional distribution $\mathcal{D}_{\sigma|Y}$ (except in one application where $n = 7$ and the computation of the exact distribution is feasible). See the appendix in Kennedy (1995) for a discussion of some practical considerations concerning the number of permutations.

Instead of drawing a large number of permutations P, there also exists another approximation to the conditional distribution: its limiting counterpart. Thus, we can employ the conditional asymptotic distribution $\mathcal{D}_{\infty|Y}$ which is obtained from $\mathcal{D}_{\sigma|Y}$ for $n \to \infty$. For the supLM test, the joint asymptotic conditional distribution of the vector of standardized statistics $Z = (Z_{\pi_1}, \ldots, Z_{\pi_m})^\top$ is multivariate normal. Therefore, it is relatively easy to compute $\mathcal{D}_{\infty|Y}$ because efficient numerical algorithms are available for computing p values for the maximum of a multivariate normal statistic Z (Genz 1992). Thus, it is computationally cheap (for small to moderate m) to compute the asymptotic conditional distribution $\mathcal{D}_{\infty|Y}$ while the advantage of the somewhat more costly computation of $\mathcal{D}_{\sigma|Y}$ is that the quality of this approximation can be controlled by choosing a sufficiently large P. In a permutation context, it seems more natural to treat the breakpoints π as fixed, however, it is also possible to let the number of breakpoints π grow with n (see Hothorn and Zeileis 2008, for theoretical and practical consequences).

The asymptotic normality of Z stated in the previous paragraph still needs to be stated more precisely and, of course, proved. It can be shown that expection, variance and covariance of Z under H_0 and given all permutations $\sigma \in S$ is:

$$\begin{aligned}
\mathsf{E}_\sigma[Z_\pi] &= 0 \quad (\pi \in \Pi) \\
\mathsf{VAR}_\sigma[Z_\pi] &= 1 \quad (\pi \in \Pi) \\
\mathsf{COV}_\sigma[Z_\pi, Z_\tau] &= \frac{n_{1,\pi} n_{2,\tau}}{\sqrt{n_{1,\pi} n_{2,\pi} n_{1,\tau} n_{2,\tau}}} \quad (\pi < \tau)
\end{aligned}$$

Collecting the variances and covariances in the matrix Σ, the multivariate normality of Z can be compactly stated as $Z \sim \mathcal{N}(0, \Sigma)$. A formal proof is given in the appendix which is obtained by embedding the test statistics Z_π and $\max_{\pi \in \Pi} |Z_\pi|$ into the framework of Strasser and Weber (1999) who establish asymptotic normality for a general class of permutation tests. This is also the reason for using $n - 1$ rather than n in the standardization of Z_π. Here, we follow the formulation of Strasser and Weber (1999), whereas for F_π we use the standard n in the LM statistic. Note that this only influences the p values computed from the two asymptotic distributions $\mathcal{D}_\infty$ and $\mathcal{D}_{\infty|Y}$, whereas the p values from $\mathcal{D}_{\sigma|Y}$ remain unaffected. Furthermore, the difference in standardization only has an influence for small n and will lead to slightly smaller p values for the unconditional asymptotic distribution $\mathcal{D}_\infty$ (but as we will see below, this does not make any difference in practice).

The assumptions under which the unconditional and conditional distributions are valid reference distributions for the test statistic are as different as the underlying conceptual frameworks. The unconditional asymptotics can be established under different sets of assumptions

such as those given in Andrews (1993) or Krämer, Ploberger, and Alt (1988), typically requiring some weak dependence of the series Y_i $(i = 1, \ldots, n)$ and certain regularity assumptions for the estimators employed. The assumptions for the conditional permutation tests, on the other hand, are simpler but in a time series setup somewhat more restrictive: they require exchangeability of the observations Y_i under H_0, i.e., the joint distribution of $(Y_1, \ldots, Y_n)$ is required to be invariant with respect to the group of permutations S. In the model-based view, this is equivalent to exchangeability of the errors ε_i, respectively. The issues are also discussed in Remark 2.4 of Strasser and Weber (1999) and the discussion in Kennedy (1995).

Finally, extreme value asymptotics can be employed to obtain the limiting distribution when no trimming is applied and all conceivable changepoints are considered. More precisely, Antoch and Hušková (2001) show that the untrimmed test statistic $\sup_{\pi \in (0,1)} |Z_\pi|$ – i.e., with $\Pi = (0, 1)$ – has an extreme value distribution that is denoted $\mathcal{D}^0_{\infty|Y}$ in the following. Analogously to the ideas above, the corresponding exact conditional distribution $\mathcal{D}^0_{\sigma|Y}$ can again be approximated by P draws from the permutation distribution of the untrimmed statistic.

2.3. Finite sample performance

To illustrate the quality of the reference distributions D for the test statistic in scenarios with small sample size n, a Monte Carlo study of a local alternative model is conducted:

$$Y_i \quad = \quad 0 \; + \; n^{-1/2}\delta \cdot \mathbf{1}_{(t^*,1]}(t_i) \; + \; \varepsilon_i$$

where $\mathbf{1}_I$ is the indicator function for the interval I, δ controls the intensity and t^* the timing of the shift. Thus, the mean of Y_i jumps from 0 to $n^{-1/2}\delta$ after time t^*. The standardized time is simply $t_i = i/n$ and the disturbances ε_i are standard normal and independent.

To study the influence of the various parameters of the model, the number of observations n is set to 10, 20 and 50, respectively, $t^* = 0.2, 0.35, 0.5$ and $\delta = 0, 5, 10, 15$. The earliest shift is $t^* = 0.2$ so that for the smallest sample size $n = 10$ there are two observations in the first regime. For comparing the performance of the distributions D, power curves (at significance level 5%) are estimated from 10,000 replications for each parameter combination. The values for the shift intensity δ also include 0 to analyze size as well as power of the tests. This setup corresponds to power/size "conditional on assignment" (in the terminology of Kennedy 1995) allowing for a fair comparison between the unconditional and conditional version of the supLM test.

The results from the Monte Carlo experiment are summarized both in Table 1 and Figure 1. These clearly indicate that the (approximated) exact conditional distribution $\mathcal{D}_{\sigma|Y}$ performs best, both in terms of power and size, independent of the timing of the shift. Among the asymptotic distributions, the conditional asymptotic distribution $\mathcal{D}_{\infty|Y}$ outperforms the unconditional asymptotic distribution $\mathcal{D}_\infty$. However, the differences are only large for very small sample sizes n and diminish with increasing n: for $n = 50$ the power curves are already almost indistinguishable. This justifies the usage of the unconditional limiting distribution $\mathcal{D}_\infty$ (typically computed using the algorithm of Hansen 1997) in moderate to large samples. For small samples, however, the conditional inference approach using permutation tests for structural change proves to be a more powerful strategy.

The untrimmed tests are always outperformed by the corresponding trimmed tests. This is, of course, not surprising given that no changes were simulated to be in the trimmed

		$t^* = 0.2$				$t^* = 0.35$				$t^* = 0.5$			
n	D	$\delta = 0$	5	10	15	0	5	10	15	0	5	10	15
10	$\mathcal{D}_\infty$	0.0	0.1	2.3	15.7	0.0	0.3	4.7	25.8	0.0	0.5	7.2	35.3
	$\mathcal{D}_{\infty\|Y}$	0.6	3.4	25.5	71.0	0.6	5.6	39.5	85.5	0.7	7.9	50.8	92.0
	$\mathcal{D}^0_{\infty\|Y}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathcal{D}_{\sigma\|Y}$	3.4	13.9	50.1	85.1	3.6	25.1	77.5	97.8	3.5	35.0	90.5	99.7
	$\mathcal{D}^0_{\sigma\|Y}$	3.4	13.8	50.0	85.0	3.6	25.2	77.4	97.8	3.5	35.0	90.5	99.7
20	$\mathcal{D}_\infty$	1.4	12.7	64.2	97.4	1.2	20.6	85.1	99.8	1.4	23.6	89.5	100.0
	$\mathcal{D}_{\infty\|Y}$	3.2	20.3	76.2	99.1	2.9	30.8	92.2	100.0	2.9	35.0	94.8	100.0
	$\mathcal{D}^0_{\infty\|Y}$	0.0	0.7	13.0	61.7	0.0	1.5	30.4	87.6	0.0	2.0	36.7	91.6
	$\mathcal{D}_{\sigma\|Y}$	5.0	27.9	83.6	99.6	4.8	39.4	95.4	100.0	5.0	44.3	97.0	100.0
	$\mathcal{D}^0_{\sigma\|Y}$	4.8	27.0	82.2	99.3	4.6	38.8	95.2	100.0	4.8	43.8	96.9	100.0
50	$\mathcal{D}_\infty$	2.8	23.9	84.9	99.8	2.7	35.6	96.0	100.0	2.8	40.9	97.5	100.0
	$\mathcal{D}_{\infty\|Y}$	4.3	29.0	88.6	99.9	3.8	41.8	97.1	100.0	4.2	47.3	98.3	100.0
	$\mathcal{D}^0_{\infty\|Y}$	0.3	6.2	56.2	97.1	0.4	11.5	80.2	99.8	0.4	14.3	86.0	100.0
	$\mathcal{D}_{\sigma\|Y}$	5.2	31.4	90.0	99.9	4.4	44.6	97.5	100.0	4.9	50.0	98.6	100.0
	$\mathcal{D}^0_{\sigma\|Y}$	5.0	28.6	87.4	99.8	4.5	40.4	96.6	100.0	4.8	45.7	97.9	100.0

Table 1: Simulated power (in %) of the supLM tests.

regions. However, it is noteworthy that while the untrimmed (approximated) exact conditional distribution $\mathcal{D}^0_{\sigma|Y}$ is very close to its trimmed counterpart $\mathcal{D}_{\sigma|Y}$ (and hence not shown in Figure 1), the asymptotic conditional distribution without trimming $\mathcal{D}^0_{\infty|Y}$ performs even worse compared to its trimmed counterpart $\mathcal{D}_{\infty|Y}$. Apparently, the extreme value asymptotics require somewhat larger sample sizes to be useful as an approximation for finite sample.

In summary, this suggests that (a) the (approximated) conditional distribution is worth the computational effort in very small samples, (b) asymptotic distributions perform similarly already for moderately large samples, and (c) the conditional asymptotic distributions (keeping the number of potential changepoints fixed as $n \to \infty$) work very well for sample sizes in between. This also confirms and complements previous findings, e.g., in the context of maximally selected statistics (see Hothorn and Lausen 2003; Hothorn and Zeileis 2008).

2.4. An illustration

To illustrate the different approximations of the reference distribution $\mathcal{D}$ in an empirical application, we reanalyze a time series giving the number of youth homicides in Boston, USA. To address the problem of high youth homicide rates in Boston, a policy initiative called the "Boston Gun Project" was launched in early 1995, implementing in particular an intervention called "Operation Ceasefire" in the late spring of 1996. As a single shift alternative seems plausible but the precise start of the intervention cannot be determined, Piehl, Cooper, Braga, and Kennedy (2003) chose to model the number of youth homicides in Boston using modifications of the F tests for structural change of Andrews (1993) based on monthly data ($n = 77$ observations) from 1992(1) to 1998(5) (see Figure 2) and assessing the significance via Monte Carlo results instead of the standard reference distribution $\mathcal{D}_\infty$.

Here, we take a similar approach and test whether the number of homicides (in continuity-

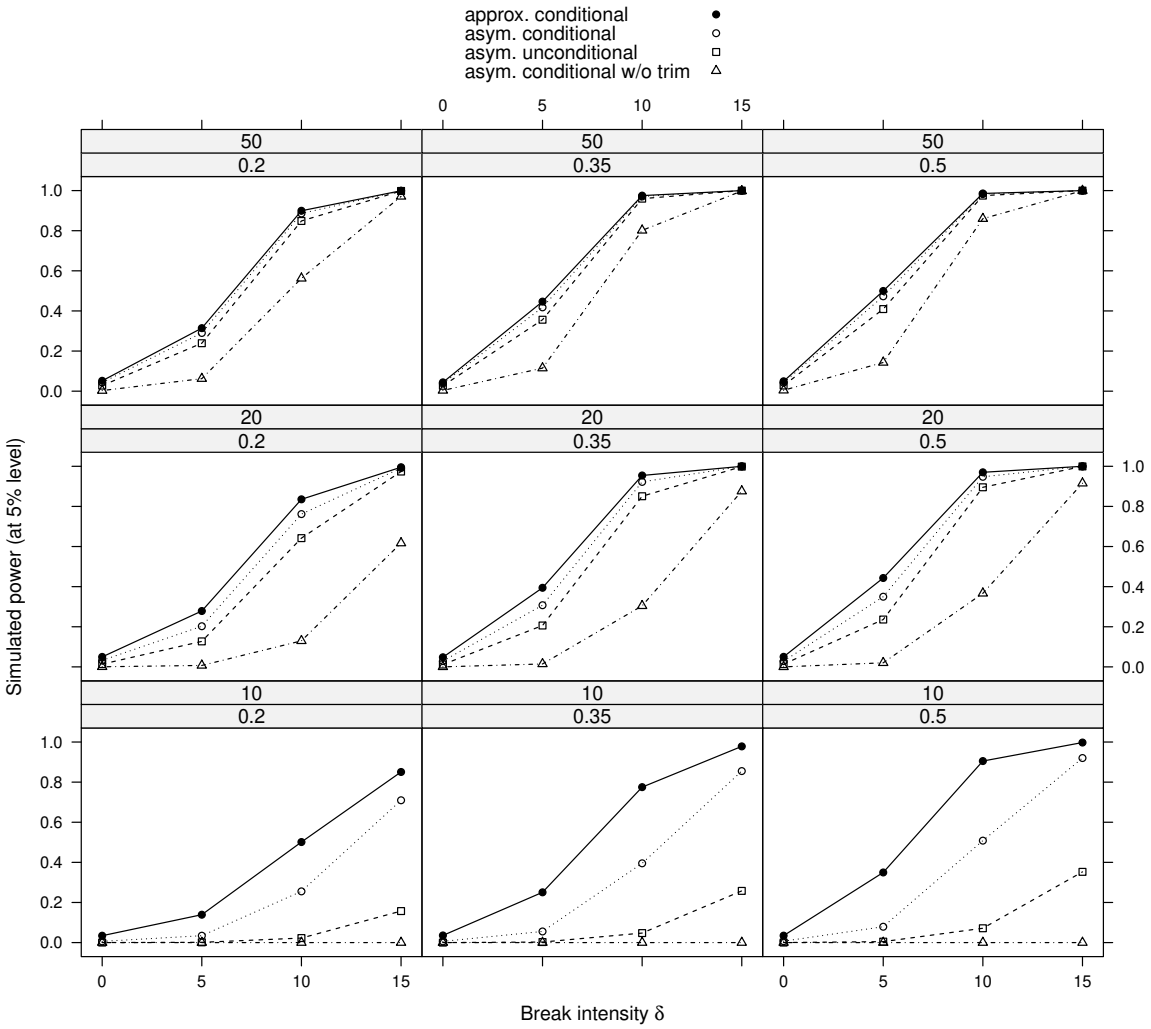

Figure 1: Simulated power of the supLM tests using reference distributions: approximated conditional $\mathcal{D}_{\sigma|Y}$ (solid, solid circles), asymptotic conditional $\mathcal{D}_{\infty|Y}$ (dotted, hollow circles), asymptotic unconditional $\mathcal{D}_{\infty}$ (dashed, hollow squares), asymptotic conditional without trimming $\mathcal{D}^0_{\infty|Y}$ (dash-dotted, hollow triangles).

corrected logarithms, $Y = \log(\text{homicides} + 0.5)$) changes over time using the supLM test of Andrews (1993) and compare the outcome of all three reference distributions $\mathcal{D}$: The test statistic is $\max_{\pi\in\Pi} |Z_\pi| = 5.374$ (or equivalently $\max_{\pi\in\Pi} F_\pi = 29.261$) with the standard asymptotic unconditional distribution $\mathcal{D}_\infty$ yielding a p value of $2.55 \cdot 10^{-6}$, the asymptotic conditional distribution $\mathcal{D}_{\infty|Y}$ a p value of $1.01 \cdot 10^{-6}$, and the approximated conditional distribution $\mathcal{D}_{\sigma|Y}$ a p value of $1 \cdot 10^{-4}$ (i.e., not a single of the 10,000 permutations produced a greater test statistic). Thus, all three p values are very similar and lead to practically equivalent solutions, providing firm evidence for a change in the number of homicides. The maximal LM statistic is assumed in 1996(7) (an estimate for the timing of the shift t^*) at about the time the Operation Ceasefire was implemented.

	1992			1993			1994			1995			1996			1997			1998		
Monthly	2	1	1	5	1	4	2	5	6	3	4	2	4	2	1	1	3	1	1	1	0
	1	3	3	2	7	3	1	1	6	1	2	3	3	4	2	0	2	3	0	2	
	3	5	4	4	5	7	3	3	3	7	3	10	3	1	2	0	1	0			
	7	2	5	2	4	4	4	3	1	5	3	3	0	1	2	1	0	3			
Annual	3.083			4.000			3.167			3.833			2.083			1.250			0.800		

Table 2: Number of youth homicides in Boston: monthly counts and annual averages.

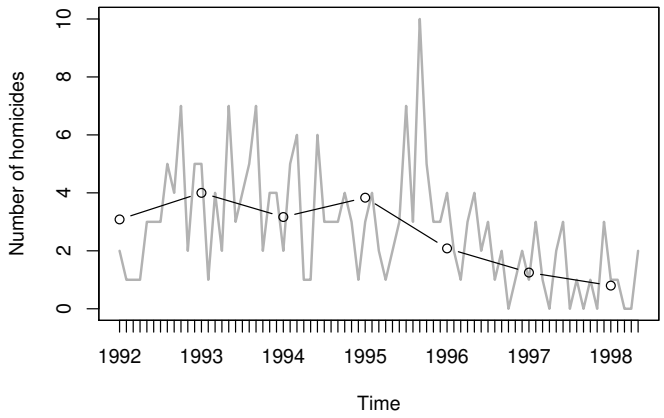

Figure 2: Number of youth homicides in Boston: monthly counts (gray), annual averages (black).

So far, we essentially confirmed the findings of Piehl *et al.* (2003) and also the impression from our simulation study that already for moderately sized n all three reference distributions $\mathcal{D}$ lead to virtually identical results. However, imagine that we would not have been provided with such detailed monthly observations but with annual averages instead (see Table 2 and Figure 2). Then, with only $n = 7$ observations, would we still be able to show that the policy intervention had an effect on the number of homicides? Using the raw means (instead of logs because we already averaged) leads to a test statistic of $\max_{\pi\in\Pi} |Z_\pi| = 2.246$ (or equivalently $\max_{\pi\in\Pi} F_\pi = 5.885$). This corresponds to a p value of 20.25% computed from the standard asymptotic unconditional distribution $\mathcal{D}_\infty$, 10.62% for the asymptotic conditional distribution $\mathcal{D}_{\infty|Y}$, and 5.71% for the exact conditional distribution $\mathcal{D}_{\sigma|Y}$. Thus, we observe a similar phenomenon as in the simulation study: the standard $\mathcal{D}_\infty$ lacks power and results in a clearly nonsignificant p value, whereas the conditional p values are considerably smaller. The exact p value is statistically significant at the 10% level (and on the verge of being significant at 5% level) and in fact there was not a single permutation yielding a greater test statistic, all 5.71% permutations are ties with the observed maximal test statistic (which is assumed for the year 1995).

3. Extensions

Year	1991	1992	1993	1994	1995	1996
Female hires	2	0	0	0	5	14
Male hires	427	86	104	180	111	59

Table 3: Annual hiring data of employment discrimination case.

In the previous section, we discussed how the conditional distributions $\mathcal{D}_{\sigma|Y}$ and $\mathcal{D}_{\infty|Y}$ can be established by embedding the supLM test of Andrews (1993) into the framework of Strasser and Weber (1999) for the location model (2). Here, we discuss how this general framework can be employed more generally in a structural change setup.

Strasser and Weber (1999) provide a very general approach for assessing the dependence of a sequence of possibly multivariate observations Y_i on another variable t_i by employing test statistics of the form

$$T \quad = \quad \text{vec}\left(\sum_{i=1}^{n} g(t_i)h(Y_i, (Y_1, \ldots, Y_n))^\top\right), \tag{7}$$

where $g(\cdot)$ and $h(\cdot)$ are possibly vector-valued transformations of t_i and Y_i, respectively. They are also called regression function and influence function, respectively, where the latter may depend on the full sequence of observations $(Y_1, \ldots, Y_n)$, however, only in a permutation-symmetric way. Given exchangeability of the observations, the asymptotic conditional multivariate normality of T under the null hypothesis of independence of Y_i and t_i is derived by Strasser and Weber (1999).

The choice of $g(\cdot)$ and $h(\cdot)$ determines against which types of dependence of Y_i on t_i tests based on T have good power. For a single shift alternative with unknown breakpoint as in (3), it is straightforward to use a multivariate regression function constructed from indicator functions for all potential breakpoints $g(t) = (\mathbf{1}_{[0,\pi_1]}(t), \ldots, \mathbf{1}_{[0,\pi_m]}(t))^\top$. While $g(\cdot)$ reflects the type of time dependence, the choice of h determines what types of changes in the distribution $\mathcal{F}_t$ can be captured (well): For shifts in location using the identity $h(Y) = Y$ is suitable. To aggregate the multivariate statistic T to a single scalar test statistic, typically the maximum of the standardized T is used

$$\max\left|\frac{T - \mathsf{E}_\sigma[T]}{\sqrt{\text{diag}(\mathsf{VAR}_\sigma[T])}}\right|$$

which corresponds to taking the maximum over the components of $g(\cdot)$ (i.e., the various potential breakpoints) and of $h(\cdot)$ (if it is multivariate). For the location model, more details are given in the appendix.

In the following, other choices of $h(\cdot)$ are discussed which are suitable for assessing changes in binary observations Y_i, multivariate series, stratified data (including certain types of panel data) and parametric models, respectively. In all illustrations, the exact conditional distribution $\mathcal{D}_{\sigma|Y}$ is used for computing p values and approximated by drawing $P = 10,000$ permutations. The section is concluded by some remarks concerning the applicability of the tests to dependent observations.

3.1. Structural changes in binary variables

For binary observations Y_i, the distribution $\mathcal{F}_{t_i}$ is binomial with a certain success probability μ_{t_i} which could depend on the time t_i. The null hypothesis of structural stability can again be

written as in (3) corresponding to constancy of the success probability. A test statistic T that compares empirical proportions from two subsamples defined by a set of potential break points $\pi_1, \ldots, \pi_m$ can simply be obtained by using a dummy coding for Y_i. This corresponds to using the influence function $h(Y) = \mathbf{1}_{\{\text{success}\}}(Y)$ while the remaining ingredients of the test remain the same and can be applied out of the box. As an illustration we use the data provided in Table 3 from an employment discrimination case described in Freidlin and Gastwirth (2000). The issue in the case was whether the hiring policy was gender neutral and a charge was filed in May 1994. Freidlin and Gastwirth (2000) supported the court's decision that there was evidence that the employer switched from under-hiring of females (compared to a fraction of 3.43% in the qualified labor force in the labor market) to over-hiring after the charge was filed. Employers can use such strategies to obscure discriminations in data aggregated over time (using both pre- and post-charge periods). Here, we reanalyze the data set in a simple structural change setup, i.e., without employing the additional knowledge of the fraction of females in the qualified labor force. Using the test procedure described above, we show that the fraction of hired females changed significantly over the years. Although there are $n = 988$ observations, there are only $m = 5$ potential breakpoints (in $1991, \ldots, 1995$) as the data is reported annually. Thus, the regression function is $g(t) = (\mathbf{1}_{[1991,1991]}(t), \ldots, \mathbf{1}_{[1991,1995]}(t))^\top$ and $h(Y) = \mathbf{1}_{\{\text{female}\}}(Y)$ yielding a 5-dimensional statistic T. The maximum of the standardized statistics is 10.49 assumed in 1995, corresponding to a p value of 10^{-4} (i.e., not a single permutation yielded a greater test statistic) conforming with the findings of Freidlin and Gastwirth (2000).

3.2. Structural changes in multivariate series

If the observations Y_i are vector-valued, several scenarios are conceivable: all components correspond to dependent variables, or some might also correspond to independent variables, or there might be one stratifying variable. The latter two scenarios are dealt with in the next paragraphs, here we focus on the case of a multivariate dependent series of observations Y_i. Typically, a multivariate influence function $h(\cdot)$ is used which is obtained by applying a suitable univariate influence function to each component of Y_i. Thus, a sequence of standardized statistics (over potential breakpoints) is computed for each component and the test rejects the null hypothesis of stability if there is evidence for structural change in any of the components. By using the joint distribution of all standardized statistics, this procedure corrects appropriately for multiple testing via incorporation of the full correlation structure over time and components. This allows not only for identification of the timing of the shift (as in the previous illustrations) but also of the component of Y_i affected by it.

To demonstrate this approach in an empirical application, the following environmetric task is considered: Coarse woody debris (CWD, dead wood ≥ 10 cm diameter) is a large stock of carbon in tropical forests, yet the flux of carbon out of this pool, via respiration, is poorly resolved (Chambers, Schimel, and Nobre 2001). The heterotrophic process involved in CWD respiration should respond to reductions in moisture availability, which occurs during dry season (Chambers *et al.* 2001). CWD respiration measurements were taken in a tropical forest in west French Guiana, which experiences extreme contrasts in wet and dry season (Bonal *et al.* 2008). An infrared gas analyzer and a clear chamber sealed to the wood surface were used to measure the flux of carbon out of the wood (Stahl, Burban, Goret, and Bonal 2011). Measurements were repeated on six pieces of wood 13 times from July to November 2011, during the transition into and out of the dry season. The aim is to assess if there

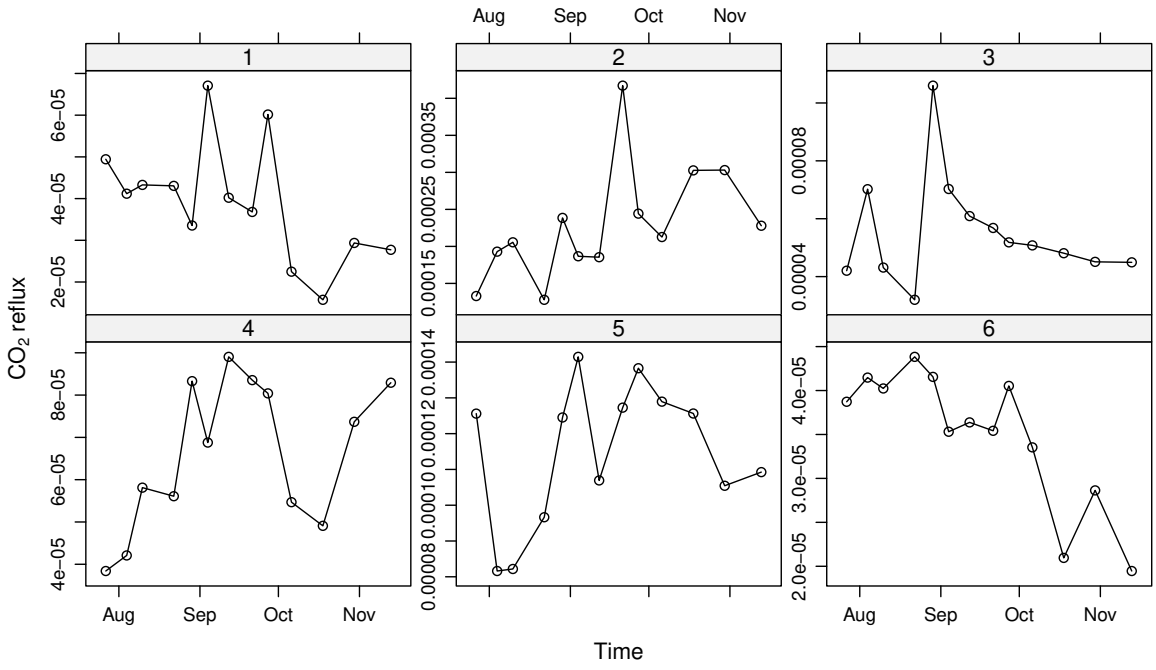

Figure 3: CO_2 reflux in coarse woody debris for eight pieces of wood.

were shifts in the CWD respiration of any of the pieces in response to the transition into (early August) and out of (late October) the dry season. The six time series are displayed in Figure 3.

We investigate the six-variate series of CO_2 reflux aiming to find out whether the CO_2 reflux has changed over sampling period in at least one of the six wood pieces. As the observations Y_i are numeric, we employ the identity transformation as the influence function $h(Y) = Y$ (which is consequently six-variate here) and the regression function $g(t)$ corresponds to $m = 10$ potential breakpoints (with a trimming of 10%) yielding a statistic T of dimension 10×6. The maximum of the standardized statistics is 3.08 corresponding to a p value of $83.9 \cdot 10^{-4}$. This maximal statistic is assumed on 2011-10-06 for the eigthth series. Only the statistics for the sixth wood sample exceeds the 95% critical value of 2.86, thus signalling a significant change in that piece of wood only (see also Figure 4).

3.3. Structural changes in stratified observations

If one of the components of a multivariate sequence Y_i stratifies the observations into *independent* blocks, the following simple strategy can be used: Compute the statistic T and its association expectation and covariance (given $\sigma \in S$) for each block and aggregate the block-wise statistics by taking their sum. Due to independence of the blocks the expectation and covariance of the aggregated statistic is also obtained by taking sums. In addition to independence the (hypothesized) block-wise breakpoints should be identical (or at least similar) for the test to have good power. In econometric applications, such situations occur less often compared to planned experiments in statistical applications—however, some situations are

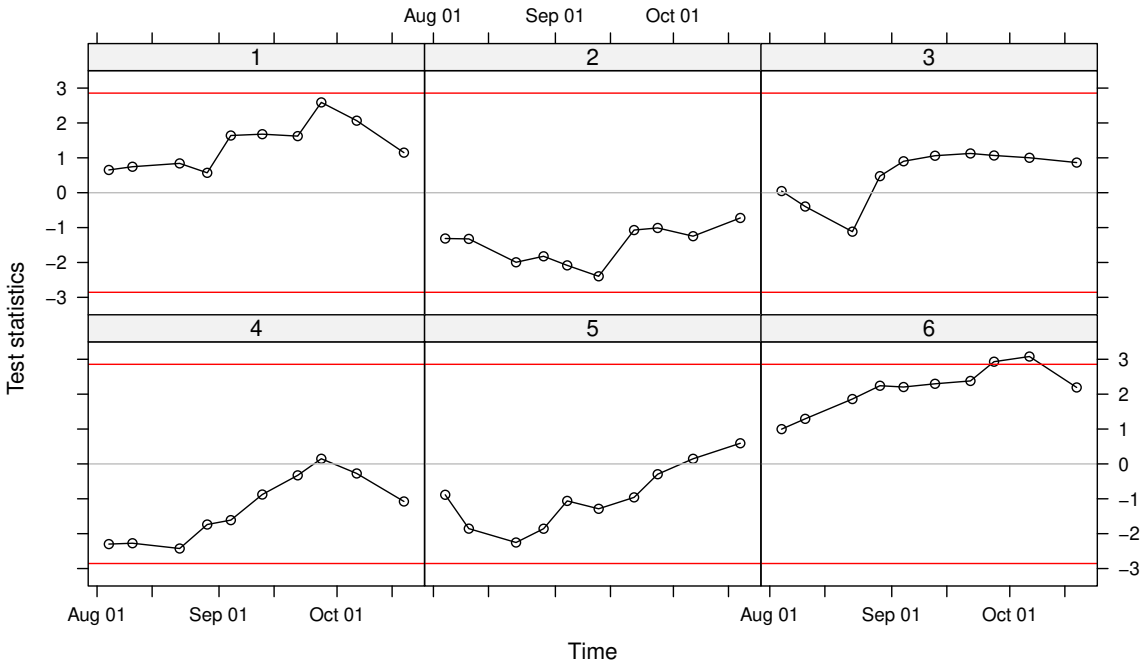

Figure 4: Test statistics for carbon reflux data.

conceivable (e.g., panel data from independent companies). Another application would be to use a multivariate influence function $h(\cdot)$ and treat its components as blocks. This is useful for model-based tests (see below) when decorrelated score functions for different parameters are used for $h(\cdot)$.

3.4. Structural changes in parametric models

To assess changes in certain aspects of the distribution $\mathcal{F}_t$ a parametric model could be useful, in particular if the observations can be split up into dependent and explanatory variables $Y_i = (y_i, x_i)^\top$. As the influence function may depend on the full set of observations (in a permutation symmetric way), the model and its corresponding parameter estimate can be easily incorporated into $h(\cdot)$. As the most important special case, we first consider some options for the linear regression model

$$y_i \quad = \quad x_i^\top \theta \; + \; \varepsilon_i, \tag{8}$$

where the assumption of exchangeability now has to be fulfilled for the disturbances ε_i (under the null hypothesis) to render the permutation approach valid. Then, structural changes in the conditional mean of the y_i can be easily assessed by using the usual ordinary least squares (OLS) residuals in the influence function $h(Y_i) = \hat{\varepsilon}_i = y_i - \hat{y}_i$. For univariate observations $Y_i = y_i$ this is equivalent to the supLM test described in Section 2. Analogously, changes in the variance of the disturbances can be captured by basing the test on the squared residuals $h(Y_i) = \hat{\varepsilon}_i^2$. Moreover, changes in any component of the vector of regression coefficients θ can be tested by using the full OLS model scores $h(Y_i) = \hat{\varepsilon}_i x_i$.

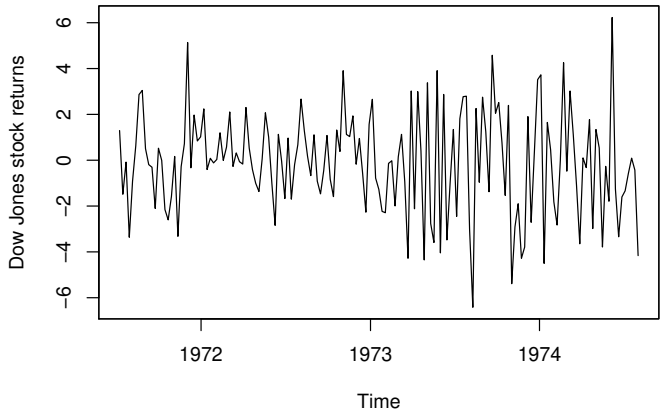

Figure 5: Dow Jones industrial average returns (in %).

Similar to the vector-valued influence function for multivariate series, this leads again to a statistic for each combination of potential breakpoint and parameter component. The same ideas apply not only to linear regression models, but to more general parametric models as long as the exchangeability assumption can be assured. Adopting a permutation approach for $h((y_i, x_i)^\top)$ is similar in spirit to the fixed regressor bootstrap of Hansen (2000), in the sense that it also permutes fixed couples of dependent and regressor variables.

More formally, we could consider a model for univariate or multivariate observations $Y \sim \mathcal{F}$ (under the null hypothesis) that are modelled by a parametric distribution $\mathcal{G}(\theta)$ (which might or might not include the true distribution $\mathcal{F}$). Then, some model scores or moment conditions derived from the model $\mathcal{G}(\theta)$, $\psi_\theta(Y) = const$ say, could be used for testing the model stability via the influence function $h(Y) = \psi_{\hat{\theta}}(Y)$. This uses the same ideas as score-based fluctuation tests, see Zeileis (2005) for a unified approach discussing different score functions $\psi_\theta(Y)$. Here, we use these ideas to investigate the stability of the distribution of Dow Jones industrial average stock returns based on weekly closing prices from 1971-07-02 to 1974-08-02. The series of prices is provided in Hsu (1979) and the corresponding log-difference returns ($\times 100$) are depicted in Figure 5. Following Hsu (1979), we model the returns Y_i as approximately normally distributed with mean and variance $\theta = (\mu, \sigma^2)^\top$ leading to the maximum likelihood scores $\psi_\theta(Y) = (Y - \mu, (Y - \mu)^2 - \sigma_2)^\top$ corresponding to the usual moment conditions. For assessing the stability of both mean and variance of the returns, we employ the empirical model scores as the influence function $h(Y) = \psi_{\hat{\theta}}(Y)$ (note that using the simpler $h(Y) = (Y, (Y - \hat{\mu})^2)^\top$ would lead to identical results) and again a 10% trimming for deriving $g(t)$. This yields a maximal standardized statistic of 4.96 assumed on 1973-03-16 for the variance, corresponding to a p value of $2 \cdot 10^{-4}$—the maximal statistic for the mean, on the other hand, is considerably smaller with 1.89, remaining clearly below the 95% critical value of 3.19. Thus, there is evidence for a clear shift in the variances in mid-March 1973 (matching the break found by Hsu 1979) while the mean remains constant throughout the sample period.

Note that exchangeability does not appear to be too strong an assumption for this data,

although one might expect GARCH (generalized autoregressive conditional heteroskedsticity) effects to be present. However, the first order autocorrelations of both the levels and its squares are very modest: Up to mid-March 1973 the autocorrelation in the levels is 0.220 and only -0.144 in the squares. After the detected break, the autocorrelations are -0.158 and -0.072, respectively. None of these would be picked up as significant by Box-type tests for autocorrelation.

3.5. Recursive partitioning of cross-section data

While the assumption of exchangeability may be questionable for time series data – as it may (illustrated in Section 3.4) or may not hold (see Section 3.6 for a brief discussion) – it is typically far less critical for cross-section data. And although one might think that there are not many structural change questions for cross-section data, this is in fact not the case.

There is a close connection between testing/estimating structural changes and estimating regression/classification trees (also known as recursive partitioning, see e.g., Hastie, Tibshirani, and Friedman 2009 for an overview). To capture the relationship between some dependent variable Y_i and a set of ℓ regressors $t_i^{(1)}, \ldots, t_i^{(\ell)}$, trees recursively assess whether the distribution of Y_i changes across any of the $t_i^{(j)}$. If there is some change, it is captured by splitting the data, i.e., estimating a breakpoint (or several breaks) with respect to the most relevant variable $t_i^{(j)}$, and then recursively repeating the procedure. While traditional tree methods solve this problem by greedy forward search algorithms, modern tree methods employ significance tests for determining in each step which variable is split next. In the latter class of tree methods, some authors employ general association tests (e.g., Hothorn, Hornik, and Zeileis 2006b), others use maximally-selected statistics (e.g., Strobl, Boulesteix, and Augustin 2007), but also structural change tests have already been suggested (Zeileis, Hothorn, and Hornik 2008).

Here, we note that the framework of maximally-selected permutation tests for structural change – as discussed in this manuscript – is a natural candidate for the basis of recursive partitioning methods. In fact, this is a special case of the conditional inference trees framework suggested by Hothorn *et al.* (2006b) with appropriately chosen influence and regression functions $h(\cdot)$ and $g(\cdot)$, respectively. In particular, the vector of indicators for all conceivable splits in a particular variable $t_i^{(j)}$ can again be used for the regression function $g(\cdot)$, whereas $h(\cdot)$ will again depend on the type of response as discussed above.

To demonstrate the flexibility of this approach is, we recursively partition a linear regression model $y_i = x_i^\top \theta + \varepsilon_i$ using the associated OLS model scores $h(Y_i) = \hat{\varepsilon} x_i$ as the influence function (see Section 3.4). For illustration, we reanalyze economic data previously considered by Zeileis *et al.* (2008), but instead of using their fully parametric model-based recursive partitiong approach (based on supLM tests), we employ a semi-parametric approach (based on permutation tests).

The data set considered is related to journal pricing, a topic that has received considerable attention in the economics literature in recent years, see Bergstrom (2001) and his journal pricing web page `http://www.econ.ucsb.edu/~tedb/Journals/jpricing.html` for further information. Using data collected by T. Bergstrom for $n = 180$ economics journals, Stock and Watson (2007) fit a demand equation by OLS for the number of library subscriptions explained by the price per citation (both in logs). Their analysis suggests price elasticity depends on

Subsample	Regression variables		Partitioning variables			
	Intercept	Slope	Society	Pages	Characters	Age
All	4.766	−0.533	1.118	2.958	2.426	5.736
($n = 180$)	0.056	0.036	0.892	0.291	0.747	< 0.001
Age ≤ 18	4.353	−0.605	0.295	1.228	1.216	2.252
($n = 53$)	0.117	0.075	0.999	1.000	1.000	0.605
Age > 18	5.011	−0.403	0.745	2.254	1.510	2.149
($n = 127$)	0.060	0.038	0.991	0.884	1.000	0.884

Table 4: Recursive partitioning of journals data. For the regression, the coefficient estimates and corresponding standard errors are provided. For the partitioning variables, the maximum test statistics with associated Bonferroni-corrected p values are provided.

the age but it is not clear in which form. Zeileis *et al.* (2008) show in their analysis that using separate elasticities for young and old journals, respectively, captures the relationship very well. Here, we reconsider their analysis using permutation tests. Thus, we use a linear regression model for $y_i = \log(subscriptions_i)$ explained by $x_i = (1, \log(price_i / citations_i))^\top$ using $\ell = 4$ variables $t_i^{(j)}$ for partitioning: a binary indicator whether the journal is published by a *society* or not, the number of *pages*, the *characters* per page, and the *age* of the journal (in years in 2000).

Table 4 shows the fitted regression coefficients along with their associated standard errors in the middle column and the results of the permutation tests (statistic and Bonferroni-adjusted p value) for all four partitioning variables. First, the model is estimated for all observations, leading to a rather low price-elasticity of the demand for journals of only −0.533. Then the model's stability is assessed through permutation tests for structural change, showing a highly significant instability with respect to the journal's age. Minimizing the residual sum of squares of a linear regression with a single break leads to a split at an age of 18 years. For the older more established journals the price elasticity is even lower than in the full sample (−0.605) while for the younger journals the price elasticity is somewhat larger (−0.403). No significant instabilities remain as all the structural change tests in the two subsamples are nonsignificant and hence the recursive partitioning stops. The fitted partitioned model is also visualized in Figure 6.

To gain some more insight into the partitioning of the full sample, Figure 7 shows the sequence of standardized Z statistics with respect to all possible splits in age. This yields a maximal standardized statistic of 5.74 assumed at an age of 17 for the intercept and of 4.79 assumed at an age of 11 for the slope, both clearly exceeding the 95% critical value of 3.43 (which was Bonferroni-adjusted to account for testing along four different partitioning variables). Note that the maximal statistic occurs very closely to the best breakpoint (in the sense of minimizing the residual sum of squares).

In summary, the partitioned model is exactly the same as the one estimated by Zeileis *et al.* (2008) but we confirmed that no further splits in the model are necessary. Due to the moderately large subsample sizes, the permutation test approaches can be expected to be more powerful but still all p values are clearly nonsignificant.

3.6. Dependent observations

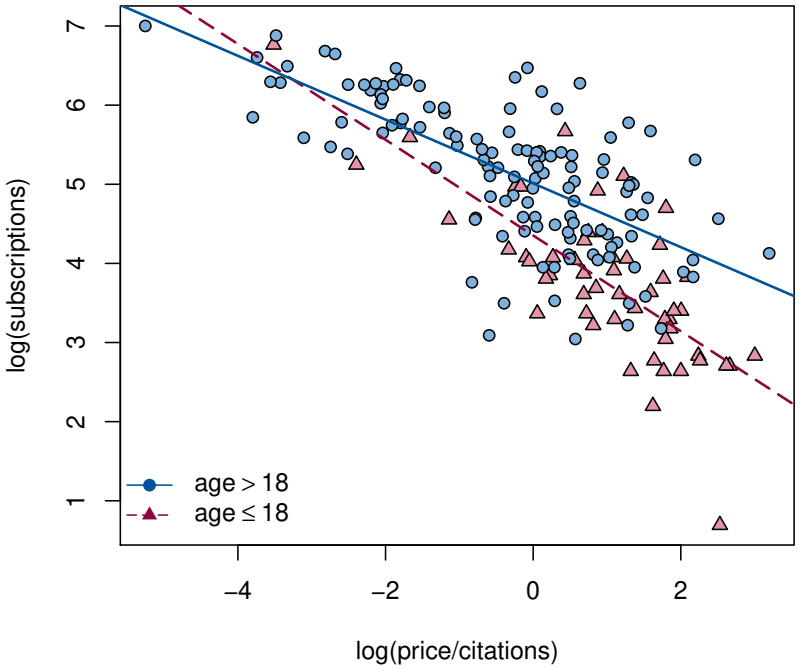

Figure 6: Recursively partitioned journals data. Blue circles represent the older journals (age > 18) with a lower price elasticity of -0.403 (blue solid line) and red triangles are the younger journals (age ≤ 18) with a higher price elasticity of -0.605 (red dashed line).

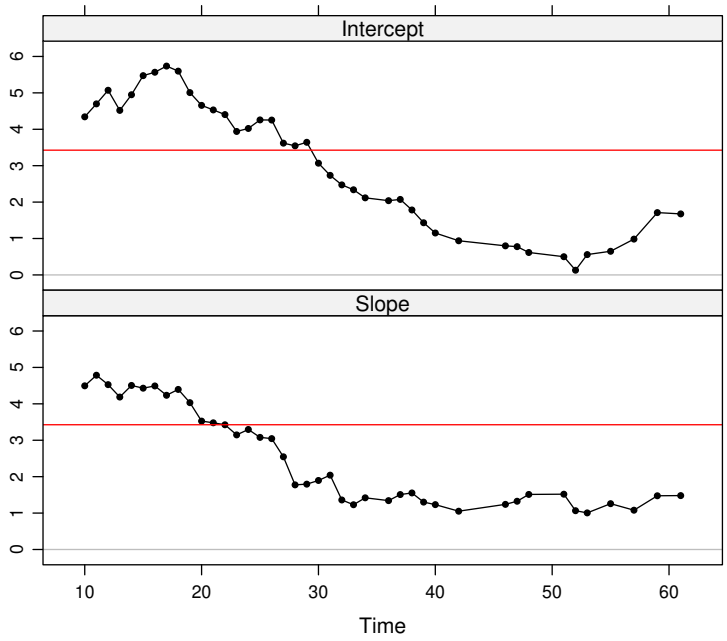

Figure 7: Test statistics for journals data with Bonferroni-corrected 5% critical value.

A crucial assumption of the permutation test approach is the exchangeability of the observations (or the disturbances in the model-based view) which is typically violated with dependent observations (Kennedy 1995; McCullagh 2005; Luger 2006). If such a violation of the assumptions is ignored, the permutation test will become liberal and not keep its size (see also Table 5 in the appendix). This is also a well-known problem for the unconditional test which is often addressed by either using a robust HAC estimate (Andrews 1991) in the computation of the supLM statistic or by computing the usual test statistic from residuals of an autoregressive (AR) model (instead of the original observations). The former strategy is not possible for permutation tests because the test is invariant to rescaling of the observations. However, the latter strategy—using residuals from an AR model—is possible (and can be regarded as a special type of influence function $h(\cdot)$). Simulations show that this version of the test keeps its size under autocorrelation and has somewhat higher power compared to the corresponding unconditional test (see Table 5 in the appendix). Another approach would be to adopt a different rerandomization scheme than the standard permutation procedure: Kirch and Steinebach (2006) and Kirch (2007) discuss refined permutation principles such as block permutations suitable for testing autoregressive series. Further strategies for dealing with dependent observations are available for special types of regression models (as in McCullagh 2005), but not for all models potentially of interest.

4. Conclusions

The supLM test for structural change of Andrews (1993) is embedded into a permutation test framework for the location-shift model. This yields the conditional permutation distribution of the test statistic (and its asymptotic counterpart) which can be used for inference instead of the usual unconditional asymptotic distribution. Comparing the size and power of the test procedures based on different versions of the reference distribution shows that (unconditional) asymptotics work well already for moderately large samples. In small samples, however, performance can be improved significantly by employing the conditional approach, in particular by computing/approximating the exact conditional distribution.

Permutation tests for structural change from the framework of Strasser and Weber (1999) can, in fact, not only be derived for the simple location model: The flexible class of tests considered includes both nonparametric and parametric (model-based) permutation tests. However, the results have to be taken with a grain of salt: Exchangeability of the errors might be a too strong assumption in time series applications where the dependence structure of the observations can not be fully captured within the model. Although there are time series applications where the errors are not correlated (and exchangeability is fulfilled as in the illustrations presented above), this assumption impedes the application of permutation methods to many other models of interest. However, the exchangeability assumption is less critical in cross-section data where the suggested tests are of interest as a building block in recursive partitioning methods.

Computational details

The results in this paper were obtained with R system for statistical computing (R Development Core Team 2012), version 2.15.0 using the packages **coin** 1.0-21 (Hothorn, Hornik,

van de Wiel, and Zeileis 2008) and **strucchange** 1.4-7 (Zeileis, Leisch, Hornik, and Kleiber 2002). Both, R itself and the packages, are freely available at no cost under the terms of the GNU General Public Licence (GPL) from the Comprehensive R Archive Network at http://CRAN.R-project.org/.

Acknowledgments

The coarse woody debris respiration data were kindly provided by Lucy Rowland (School of GeoSciences, University of Edinburgh).

References

Andrews DWK (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica*, **59**, 817–858.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, **61**, 821–856.

Antoch J, Hušková M (2001). "Permutation Tests in Change Point Analysis." *Statistics & Probability Letters*, **53**, 37–46.

Bergstrom TC (2001). "Free Labor for Costly Journals?" *Journal of Economic Perspectives*, **15**, 183–198.

Bonal D, Bosc A, Ponton S, Goret JY, Burban B, Gross P, Bonnefond JM, Elbers J, Longdoz B, Epron D, Guehl JM, Granier A (2008). "Impact of Severe Dry Season on Net Ecosystem Exchange in the Neotropical Rainforest of French Guiana." *Global Change Biology*, **14**(8), 1917–1933. doi:10.1111/j.1365-2486.2008.01610.x.

Boulesteix AL, Strobl C (2007). "Maximally Selected Chi-Squared Statistics and Non-Monotonic Associations: An Exact Approach Based on Two Cutpoints." *Computational Statistics & Data Analysis*, **51**(12), 6295–6306.

Brown RL, Durbin J, Evans JM (1975). "Techniques for Testing the Constancy of Regression Relationships over Time." *Journal of the Royal Statistical Society B*, **37**, 149–163.

Chambers JQ, Schimel JP, Nobre AD (2001). "Respiration from Coarse Wood Litter in Central Amazon Forests." *Biogeochemistry*, **52**(2), 115–131.

Chow GC (1960). "Tests of Equality between Sets of Coefficients in Two Linear Regressions." *Econometrica*, **28**, 591–605.

Ernst MD (2004). "Permutation Methods: A Basis for Exact Inference." *Statistical Science*, **19**(4), 676–685.

Fisher RA (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK.

Freidlin B, Gastwirth JL (2000). "Changepoint Tests Designed for the Analysis of Hiring Data Arising in Employment Discrimination Cases." *Journal of Business & Economic Statistics*, **18**(3), 315–322.

Genz A (1992). "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics*, **1**, 141–149.

Hansen BE (1997). "Approximate Asymptotic p Values for Structural-Change Tests." *Journal of Business & Economic Statistics*, **15**, 60–67.

Hansen BE (2000). "Testing for Structural Change in Conditional Models." *Journal of Econometrics*, **97**, 93–115.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning*. Springer-Verlag, New York, 2nd edition.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2006a). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263. doi:10.1198/000313006X118430.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**(8), 1–23. URL http://www.jstatsoft.org/v28/i08/.

Hothorn T, Hornik K, Zeileis A (2006b). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933.

Hothorn T, Lausen B (2003). "On the Exact Distribution of Maximally Selected Rank Statistics." *Computational Statistics & Data Analysis*, **43**(2), 121–137.

Hothorn T, Zeileis A (2008). "Generalized Maximally Selected Statistics." *Biometrics*, **64**, 1263–1269. doi:10.1111/j.1541-0420.2008.00995.x.

Hsu DA (1979). "Detecting Shifts of Parameter in Gamma Sequences with Applications to Stock Price and Air Traffic Flow Analysis." *Journal of the American Statistical Association*, **74**, 31–40.

Kennedy PE (1995). "Randomization Tests in Econometrics." *Journal of Business & Economic Statistics*, **13**(1), 85–94.

Kirch C (2007). "Block Permutation Principles for the Change Analysis of Dependent Data." *Journal of Statistical Planning and Inference*, **137**, 2453–2474.

Kirch C, Steinebach J (2006). "Permutation Principles for the Change Analysis of Stochastic Processes Under Strong Invariance." *Journal of Computational and Applied Mathematics*, **186**(1), 64–88.

Krämer W, Ploberger W, Alt R (1988). "Testing for Structural Change in Dynamic Models." *Econometrica*, **56**(6), 1355–1369.

Krämer W, Sonnberger H (1986). *The Linear Regression Model under Test*. Physica-Verlag, Heidelberg.

Lausen B, Schumacher M (1992). "Maximally Selected Rank Statistics." *Biometrics*, **48**, 73–85.

Ludbrook J, Dudley H (1998). "Why Permutation Tests are Superior to t and F Tests in Biomedical Research." *The American Statistician*, **52**(2), 127–132.

Luger R (2006). "Exact Permutation Tests for Non-Nested Non-Linear Regression Models." *Journal of Econometrics*, **133**, 513–529.

McCullagh P (2005). "Exchangeability and Regression Models." In AC Davison, Y Dodge, N Wermuth (eds.), "Celebrating Statistics – Papers in Honour of Sir David Cox on his 80th Birthday," pp. 89–115. Oxford University Press, Oxford. doi:10.1093/acprof:oso/9780198566540.003.0005.

Pesarin F (2001). *Multivariate Permutation Tests: With Applications to Biostatistics*. John Wiley & Sons, Chichester, UK.

Piehl AM, Cooper SJ, Braga AA, Kennedy DM (2003). "Testing for Structural Breaks in the Evaluation of Programs." *Review of Economics and Statistics*, **85**(3), 550–558.

Pitman EJG (1938). "Significance Tests which May Be Applied to Samples from any Populations: III. The Analysis of Variance Test." *Biometrika*, **29**, 322–335.

Quandt RE (1960). "Tests of the Hypothesis That a Linear Regression Obeys Two Separate Regimes." *Journal of the American Statistical Association*, **55**, 324–330.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Stahl C, Burban B, Goret JY, Bonal D (2011). "Seasonal Variations in Stem CO_2 Efflux in the Neotropical Rainforest of French Guiana." *Annals of Forest Science*, **68**(4), 771–782. doi:10.1007/s13595-011-0074-2.

Stock JH, Watson MW (1996). "Evidence on Structural Instability in Macroeconomic Time Series Relations." *Journal of Business & Economic Statistics*, **14**, 11–30.

Stock JH, Watson MW (2007). *Introduction to Econometrics*. Addison-Wesley, Reading, MA, 2nd edition.

Strasser H, Weber C (1999). "On the Asymptotic Theory of Permutation Statistics." *Mathematical Methods of Statistics*, **8**, 220–250. Preprint available from http://epub.wu.ac.at/102/.

Strobl C, Boulesteix AL, Augustin T (2007). "Unbiased Split Selection for Classification Trees Based on the Gini Index." *Computational Statistics & Data Analysis*, **52**, 483–501.

Zeileis A (2005). "A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals." *Econometric Reviews*, **24**(4), 445–466. doi:10.1080/07474930500406053.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514. doi:10.1198/106186008X319331.

Zeileis A, Leisch F, Hornik K, Kleiber C (2002). "**strucchange**: An R Package for Testing for Structural Change in Linear Regression Models." *Journal of Statistical Software*, **7**(2), 1–38. URL http://www.jstatsoft.org/v07/i02/.

A. Proofs

The (asymptotic) distribution of the multivariate statistic $Z = (Z_{\pi_1}, \ldots, Z_{\pi_m})^\top$ is derived by embedding the statistic into the framework of Strasser and Weber (1999), as discussed in Hothorn *et al.* (2006a). More precisely, the test statistic $\max_{\pi \in \Pi} Z_\pi$ considered above is the maximum-type statistic $c_{\max}$ of Hothorn *et al.* (2006a) if the influence function $h(Y) = Y$ is used for the observations Y_i and the transformation $g(t) = (\mathbf{1}_{[0,\pi_1]}(t), \ldots, \mathbf{1}_{[0,\pi_m]}(t))^\top$ is used for the associated timings t_i. The transformation g used the indicator function $\mathbf{1}_I$ of the interval I and thus corresponds to a vector of indicators for the time up to the timings π_j $(j = 1, \ldots, k)$.

Using these transformations $h(\cdot)$ and $g(\cdot)$, the unstandardized test statistic T is in the notation of Hothorn *et al.* (2006a)

$$T \quad = \quad \mathrm{vec}\left(\sum_{i=1}^{n} g(t_i) h(Y_i)^\top\right) \quad = \quad (n_{1,\pi_1} \bar{Y}_{1,\pi_1}, \ldots, n_{1,\pi_m} \bar{Y}_{1,\pi_m})^\top. \tag{9}$$

Under H_0, given all permutations $\sigma \in S$ of the observations $Y_1, \ldots, Y_n$, the unstandardized statistic has expectation

$$\mathsf{E}_\sigma[T] \quad = \quad \mathrm{vec}\left(\left(\sum_{i=1}^{n} g(t_i)\right) n^{-1} \sum_{i=1}^{n} h(Y_i)^\top\right) \quad = \quad (n_{1,\pi_1}, \ldots, n_{1,\pi_m})^\top \bar{Y} \tag{10}$$

and each unstandardized statistic has variance

$$\mathsf{VAR}_\sigma[T_\pi] \quad = \quad \left(n_{1,\pi} - \frac{n_{1,\pi}^2}{n}\right) \frac{RSS_0}{n-1} \quad = \quad \frac{n_{1,\pi} n_{2,\pi}}{n} \frac{RSS_0}{n-1}, \tag{11}$$

where the residual sum of squares is $RSS_0 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. The equations directly follow from formula (7) in Strasser and Weber (1999).

Standardizing the vector of raw statistics $T = (T_{\pi_1}, \ldots, T_{\pi_m})^\top$ by their respective mean and standard deviation yields the vector of statistics $Z = (Z_{\pi_1}, \ldots, Z_{\pi_m})^\top$:

$$\begin{aligned} Z_\pi &= \frac{T_\pi - \mathsf{E}_\sigma[T_\pi]}{\sqrt{\mathsf{VAR}_\sigma[T_\pi]}} \\ &= \frac{n_{1,\pi} \bar{Y}_{1,\pi} - n_{1,\pi} \bar{Y}}{\sqrt{\frac{n_{1,\pi} n_{2,\pi}}{n} \frac{RSS_0}{n-1}}} \\ &= \sqrt{\frac{n_{1,\pi} n_{2,\pi}}{n}} \frac{\bar{Y}_{1,\pi} - \bar{Y}_{2,\pi}}{\sqrt{RSS_0/(n-1)}}, \end{aligned}$$

because of the following simple relationship between $\bar{Y}_{1,\pi}$, $\bar{Y}_{2,\pi}$ and $\bar{Y}$:

$$\bar{Y} \quad = \quad \frac{n_{1,\pi} \bar{Y}_{1,\pi} + n_{2,\pi} \bar{Y}_{2,\pi}}{n}.$$

Consequently, Z has zero mean and unit variance given all permutations $\sigma \in S$. Similarly, the covariance between two elements of Z, Z_π and Z_τ say, is

$$\frac{RSS_0}{n-1}\left(\sum_{i=1}^{n} \mathbf{1}_{[0,\pi]}(t_i)\mathbf{1}_{[0,\tau]}(t_i)\right) - \frac{1}{n}\frac{RSS_0}{n-1}\left(\sum_{i=1}^{n} \mathbf{1}_{[0,\pi]}(t_i)\right)\left(\sum_{i=1}^{n} \mathbf{1}_{[0,\tau]}(t_i)\right) = \frac{n\min(n_{1,\pi}, n_{1,\tau}) - n_{1,\pi}n_{1,\tau}}{n}\frac{RSS_0}{n-1}.$$

Assuming that $\pi < \tau$ and using the variance computed above, the correlation is thus

$$\frac{n_{1,\pi}n_{2,\tau}}{\sqrt{n_{1,\pi}n_{2,\pi}n_{1,\tau}n_{2,\tau}}}.$$

Given that we derived the first two moments of Z by embedding the statistic into the framework of Strasser and Weber (1999), the asymptotic normality of Z follows by application of their Theorem 2.3.

B. Power and size for autocorrelated series

To study the performance of the tests with dependent data, we use a simulation setup as in Section 2.3. The only difference is that the errors are now autocorrelated with $\varrho = 0, 0.1, 0.2, 0.3, 0.5, 0.9$. The length of the time series considered is either very short ($n = 10$) or moderate ($n = 50$) and either there is no change ($\delta = 0$) or a large shift in the mean ($\delta = 15$). Five different versions of the tests are assessed: the unconditional and conditional test ($\mathcal{D}_\infty$ and $\mathcal{D}_{\sigma|Y}$, respectively) on the original data (as in Section 2.3), the unconditional test computed with a robust HAC covariance estimate and the unconditional and conditional test computed on the residuals of an AR(1) model (fitted by OLS).

Using the uncorrected tests on the original data ($\mathcal{D}_\infty$ and $\mathcal{D}_{\sigma|Y}$, respectively), it can be seen that size distortions occur for $\varrho > 0$ even if there is no change ($\delta = 0$). For ϱ up to 0.2 these are still moderate but become very large afterwards and are even more pronounced for the conditional version of the test. However, in moderately large time series ($n = 50$), the problem can be remedied by other using a HAC correction or applying the tests to the AR(1) residuals. All of the tests keep their size and have reasonable power with the conditional test having the highest power. For very short time series ($n = 10$), however, none of the tests is able to distinguish between autocorrelation and a shift in the mean (with $\delta = 15$).

Affiliation:

Achim Zeileis
Department of Statistics
Faculty of Economics and Statistics
Universität Innsbruck
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://eeecon.uibk.ac.at/~zeileis/

Torsten Hothorn
Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik
Universität Zürich
Hirschengraben 84
8001 Zürich, Switzerland
E-mail: Torsten.Hothorn@R-project.org

			Autocorrelation ϱ						
δ	n	Test	0	0.1	0.2	0.3	0.5	0.7	0.9
0	10	$\mathcal{D}_\infty$ (orig.)	0.0	0.0	0.1	0.1	0.3	0.6	1.3
		$\mathcal{D}_{\sigma\|Y}$ (orig.)	3.6	5.8	8.4	10.6	18.8	29.1	42.9
		$\mathcal{D}_\infty$ (orig. + HAC)	9.4	7.6	5.8	5.3	3.2	2.4	2.8
		$\mathcal{D}_\infty$ (AR res.)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		$\mathcal{D}_{\sigma\|Y}$ (AR res.)	1.3	1.2	1.2	1.5	1.9	2.2	3.3
0	50	$\mathcal{D}_\infty$ (orig.)	2.9	6.2	10.3	16.6	36.9	63.4	89.3
		$\mathcal{D}_{\sigma\|Y}$ (orig.)	4.8	9.3	15.0	23.0	44.6	70.6	92.2
		$\mathcal{D}_\infty$ (orig. + HAC)	2.3	1.7	1.4	1.0	0.4	0.1	0.0
		$\mathcal{D}_\infty$ (AR res.)	2.0	1.8	1.8	1.5	1.1	0.8	1.9
		$\mathcal{D}_{\sigma\|Y}$ (AR res.)	4.0	3.9	3.8	3.2	2.5	2.3	3.8
15	10	$\mathcal{D}_\infty$ (orig.)	36.3	38.1	41.1	42.4	46.0	47.9	49.1
		$\mathcal{D}_{\sigma\|Y}$ (orig.)	99.8	99.8	99.7	99.7	98.7	97.6	93.8
		$\mathcal{D}_\infty$ (orig. + HAC)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		$\mathcal{D}_\infty$ (AR res.)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		$\mathcal{D}_{\sigma\|Y}$ (AR res.)	0.8	0.7	0.8	0.8	0.9	1.3	2.4
15	50	$\mathcal{D}_\infty$ (orig.)	100.0	100.0	100.0	99.9	99.0	95.0	92.6
		$\mathcal{D}_{\sigma\|Y}$ (orig.)	100.0	100.0	100.0	99.9	99.4	96.2	94.4
		$\mathcal{D}_\infty$ (orig. + HAC)	75.9	55.4	32.6	16.2	2.3	0.2	0.0
		$\mathcal{D}_\infty$ (AR res.)	69.0	48.5	28.2	13.6	2.2	0.4	1.0
		$\mathcal{D}_{\sigma\|Y}$ (AR res.)	87.6	73.9	53.6	33.0	8.4	1.7	2.2

Table 5: Simulated power (in %) for different versions of the test in the presence of autocorrelation.